Identifying and Mitigating the Security Risks of Generative AI

Year: 2023 | Citations: 118 | Authors: Clark W. Barrett, Bradley L Boyd, Ellie Burzstein, Nicholas Carlini, Brad Chen

---

Abstract

Every major technical invention resurfaces the dual-use dilemma -- the new technology has the potential to be used for good as well as for harm. Generative AI (GenAI) techniques, such as large language models (LLMs) and diffusion models, have shown remarkable capabilities (e.g., in-context learning, code-completion, and text-to-image generation and editing). However, GenAI can be used just as well by attackers to generate new attacks and increase the velocity and efficacy of existing attacks. This paper reports the findings of a workshop held at Google (co-organized by Stanford University and the University of Wisconsin-Madison) on the dual-use dilemma posed by GenAI. This paper is not meant to be comprehensive, but is rather an attempt to synthesize some of the interesting findings from the workshop. We discuss short-term and long-term goals for the community on this topic. We hope this paper provides both a launching point for a discussion on this important topic as well as interesting problems that the research community can work to address.